Reducing dimensionality for prediction of genome-wide breeding values

Genetics Selection Evolution

BioMed Central

Research

Open Access

Reducing dimensionality for prediction of genome-wide breeding values

Trygve R Solberg*[1], Anna K Sonesson[2], John A Woolliams[1,3] and Theo HE Meuwissen[1]

Address: [1]Norwegian University of Life Sciences, Department of Animal and Aquacultural Sciences, PO Box 5003, N-1432 Ås, Norway, [2]NOFIMA Marin, PO Box 5010, N-1432 Ås, Norway and [3]Roslin Institute (Edinburgh), Roslin, Midlothian, EH25 9PS, UK

Email: Trygve R Solberg* - trygve.roger.solberg@umb.no; Anna K Sonesson - anna.sonesson@nofima.no; John A Woolliams - john.woolliams@bbsrc.ac.uk; Theo HE Meuwissen - theo.meuwissen@umb.no

* Corresponding author

Published: 18 March 2009

Genetics Selection Evolution 2009, **41**:29 doi:10.1186/1297-9686-41-29

Received: 3 February 2009
Accepted: 18 March 2009

This article is available from: http://www.gsejournal.org/content/41/1/29

Abstract

Partial least square regression (PLSR) and principal component regression (PCR) are methods designed for situations where the number of predictors is larger than the number of records. The aim was to compare the accuracy of genome-wide breeding values (EBV) produced using PLSR and PCR with a Bayesian method, 'BayesB'. Marker densities of 1, 2, 4 and 8 N_e markers/Morgan were evaluated when the effective population size (N_e) was 100. The correlation between true breeding value and estimated breeding value increased with density from 0.611 to 0.681 and 0.604 to 0.658 using PLSR and PCR respectively, with an overall advantage to PLSR of 0.016 (s.e = 0.008). Both methods gave a lower accuracy compared to the 'BayesB', for which accuracy increased from 0.690 to 0.860. PLSR and PCR appeared less responsive to increased marker density with the advantage of 'BayesB' increasing by 17% from a marker density of 1 to $8N_e/M$. PCR and PLSR showed greater bias than 'BayesB' in predicting breeding values at all densities. Although, the PLSR and PCR were computationally faster and simpler, these advantages do not outweigh the reduction in accuracy, and there is a benefit in obtaining relevant prior information from the distribution of gene effects.

Introduction

Approaches to the use of data from molecular markers in genetic evaluation for predicting breeding values have undergone considerable development as dense genome-wide marker technologies, such as high-density, high-throughput SNP chips, have become available. Currently, considerable attention is being paid to genomic selection with the approach of predicting genome-wide breeding values. Studies have demonstrated that the potential accuracies from dense molecular information are impressive, *e.g.* [[1-6], and [7]]. For example, [7] showed that it was possible to predict breeding values of unrecorded offspring using genomic selection with accuracies of 0.86 with only a small bias, for a trait with heritability 0.5, 1000 phenotypes and an effective population size of N_e = 100. Whilst in general, the accuracies of evaluation will depend on a number of factors, one issue related to implementation is the computational demand. In [7], a Bayesian approach, 'BayesB' was used, which was computationally time-consuming and required some prior assumptions to be made concerning the potential number of QTL segregating and the prior distributions for QTL and marker effects.

This increase in the scale of molecular information results in data where, typically, the number of predictors (markers) is larger than the number of records (phenotypes). This statistical problem has been considered before, and several methods based on the multivariate regression theory, such as partial least square regression, (PLSR) and principal component regression (PCR) have been used for such situations. Both these techniques reduce the dimensionality of the set of regression variables by finding a small number of linear combinations of the original predictors, but the strategy for finding the linear combinations differ between the two methods. The regression methods have found fields of application primarily in chemometrics, econometrics and social sciences *e.g.* [8,9], but there have been only very few studies using PLSR and PCR concerned with their suitability for prediction of breeding values using large-scale molecular data, *e.g.* [10,11].

Therefore, one option for reducing the computational burden of 'BayesB' and for avoiding the use of prior distribution for marker effects is to make use of the simpler and faster PLSR and PCR algorithms. However, these algorithms have not been tested sufficiently in the context of genome-wide breeding value estimation, *e.g.* against 'BayesB' results, to decide upon their desirability of use. The study tested the hypothesis that an effective evaluation using genome-wide molecular data could be obtained using regression models of reduced dimensionality. Both PLSR and PCR were compared to the 'BayesB' for their accuracy and bias in predicting genome-wide breeding values.

Methods

Population structure and genome

Population structure

The simulation model was described in detail in an earlier paper [7]. Briefly, a population with an effective population size of N_e = 100 was simulated over 1000 generations of random selection and mating with its genome subject to mutation. In generation t = 1001, the number of animals was increased to 1000 animals by factorial mating of 50 sires (i = 1–50) and 50 dams (i = 51–100) from generation 1000. The factorial mating was achieved by mating sire 1 to dams 51–70, sire 2 to dams 52–71, sire 3 to dams 53–72 and so on, and each dam had one offspring per sire. Animals in generation t = 1001 had 1000 offspring in generation t = 1002, produced by random mating among the parents in generation t = 1001. Animals in both generation t = 1001 and t = 1002 were genotyped for SNP markers.

Simulated genome

The size and structure of the genome were the same as described in [7] so that a direct comparison of the results was possible. The genome was simulated with 10 chromosomes each with a length of 100 cM each. Four density schemes of 1, 2, 4 and 8 markers/cM was evaluated, resulting in a total number of 1010, 2020, 4040 and 8080 markers across the 10 Morgan (M) genome. This would correspond to approximately 4000 to 32000 SNP markers in the Atlantic salmon (*Salmo salar*) genome, assuming a 40 M genome, or 3000 to 24000 SNP in the cattle genome, assuming a 30 M genome, respectively http://bioinfo.genopole-toulouse.prd.fr/eadgene/Wiki/IMG/pdf/EADGENE2006_02_17.pdf. However in this paper, densities will be scaled by the N_e used to generate the markers, which was N_e = 100 here, unless stated otherwise. This is because the linkage disequilibrium between markers is a function of $4N_ec$, where c is the distance between the markers and N_e represents the marker density. Thus, the densities correspond to 1, 2, 4, and $8N_e$/M and will be expressed in this way throughout the paper.

The mutation rate of the markers was assumed to be 2.5×10^{-5} per locus per meiosis and with this mutation rate, 99% of the potential markers were segregating at t = 1001. Markers with more than two alleles segregating at t = 1001 were converted to SNP as described in [7] so that the allele frequencies were as close as possible to 0.5. The typical distribution of the minor allele frequencies of the SNP markers at t = 1001 resembled a uniform distribution with an over-representation of markers with intermediate frequencies, which reflected the selection of the most informative markers that is undertaken in practice.

The potential number of QTL was kept at 100 per chromosome, distributed evenly over each chromosome (see Fig. 1). The actual number of segregating QTL at t = 1001 depended on the mutation rate which was assumed to be 2.5×10^{-3} per locus per meiosis and resulted in the number of segregating QTL being typically 5 to 6% of the

$1N_e$/M M_1-Q1-M_2-…//…-M_{100}-Q_{100}-M_{101}

$2N_e$/M M_1-M_2-Q_1-M_3-M_4-…//…-M_{199}-M_{200}-Q_{100}-M_{201}-M_{202}

$4N_e$/M M_1-M_2-M_3-M_4-Q_1-M_5-M_6-M_7-M_8-…//…-M_{397}-M_{398}-M_{399}-M_{400}-Q_{100}-M_{401}-M_{402}-M_{403}-M_{404}

$8N_e$/M M_1-M_2-M_3-M_4-M_5-M_6-M_7-M_8-Q_1-M_9-M_{10}-M_{11}-M_{12}-M_{13}-M_{14}-M_{15}-M_{16}-…//…-M_{793}-M_{794}-M_{795}-M_{796}-M_{797}-M_{798}-M_{799}-M_{800}-Q_{100}-M_{801}-M_{802}-M_{803}-M_{804}-M_{805}-M_{806}-M_{807}-M_{808}

Figure 1
Position of marker and QTL on each chromosome. M_1, M_2, ...M_x indicate the marker position, Q_1, Q_2, ...Q_{100} indicate the QTL position. The number of markers varied from $1N_e$/M (101 markers per chromosome) to $8N_e$/M (808 markers per chromosome). The number of QTL was kept constant at 100 per chromosome.

potential number with 93% biallelic. The distribution of the QTL allele frequencies of the positive QTL resembled a U-shaped distribution. The effects of a mutational allele of the QTL were sampled from the gamma distribution with the shape parameter of 1.66 and scale parameter of 0.4 [12] with an equal probability of a positive or negative effect.

The linkage disequilibrium (LD) that is generated by this population structure is described in [7]. The r-squared value increased when the marker density increased, and followed the expected value of r-squared well when allowing for mutations. Since the r-squared estimates were close to their expected values, the population was assumed to be close to a state of recombination-drift balance.

Phenotypic values

Phenotypic values for animals were first generated in generation t = 1001, and simulated as:

$\mathbf{P}_i = \mathbf{TBV}_i + \varepsilon_i$, where $\mathbf{TBV}_i$ was the true breeding value for the i'th animal and $\varepsilon \sim N(0, \sigma^2_e)$. The variance of the additive genetic effects (σ^2_a) varied somewhat from replicate to replicate, but was on average 1 (s.e = 0.118). The environmental variance (σ^2_e) was kept constant and equalled 1. Hence, the heritability varied between replicates, but was on average 0.5 (s.e = 0.026) for all 20 replicates calculated from the $1N_e/M$ scheme.

Methods for estimating breeding values

Three methods for estimating breeding values were compared on each replicated dataset: PLSR, PCR and 'BayesB'. The basic idea of PLSR and PCR is to reduce the number of predictors with a smaller number of linear combinations of the predictors, with the additional property of pair-wise independence within the set of the constructed variables. From here on, the term latent variables will be used for these combinations of predictors applied to PLSR, while the term principal components will be used for PCR. The main difference between PLSR and PCR is in the method for constructing the latent variables or principal components. PLSR maximises the amount of covariance between the standardized predictors and response for a given number of latent variables, so that the covariance between the set of latent variables and phenotypes is as high as possible. In contrast, PCR maximises the proportion of total variance among the original predictors explained by the set of principal components. The third method, 'BayesB' makes prior assumptions on the amount of genetic variance and the distribution of gene effects, and breeding values are estimated from the data points by Bayesian methods.

Principal component regression (PCR)

For PCR, the principal components associated with the largest eigenvalues of the **X**'**X** matrix were extracted and used to predict the **y** values. The following steps were performed with PCR, when fitting c principal components:

1. Marker genotype data was organised, as $p \times m$ matrix (**X**), where p is the number of phenotypic records (1000 animals in this case), m is the number of marker genotypes. Genotypes were scored as 1 for genotype AA, 0 for heterozygote (Aa or aA) and -1 for aa. Hence the size of the **X** matrix varied from 1000 × 1010 (for 1010 markers) to 1000 × 8080 (for 8080 markers), with each column containing the set of genotypes for a single marker.

2. The marker genotype matrix **X** and **y** were standardised such that each column had a mean of zero and standard deviation of 1.

3. Singular value decomposition was then performed on the **X** matrix to find the principal components, and the c first components enter as columns in the matrix **U** [13].

4. The regression coefficients were obtained as $\mathbf{b}_{PCR} = \mathbf{US}^{-1}\mathbf{U}'$, where $\mathbf{S}^{-1}$ is a diagonal matrix of the c highest singular values obtained from step 3.

5. Calculation of estimated breeding values (EBVs) was performed as explained in section 2.3.

The correlation between TBV and EBV was calculated when c = 10, 50, 100, 150, etc. components were fitted. The number of principal components that gave the highest correlation between TBVs and EBVs was used for each density.

Partial least square regression (PLSR)

In PLSR, the latent variables are constructed whilst accounting for their relationship to the data **y**, *i.e.* the latent variables are the combinations of the **X** variables that maximise the covariance with **y**. PLSR reduces the dimension of the regression $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, where **X** is a $p \times m$ design matrix, and **y** is a $p \times 1$ data vector by performing the regression $\mathbf{y} = \mathbf{Tq} + \mathbf{e}$, where **T** is a $p \times c$ vector of 'scores', **q** is a $c \times 1$ vector of 'loadings', and generally $c << m$. **T** is calculated as **XW**, where **W** is a matrix of weights. Column h of **T**, $\mathbf{t}_h$, is chosen to maximise the covariance with the data, and this is obtained by setting the corresponding weights column, $\mathbf{w}_h$, proportional to the 'deflated' **X**'**y**. The deflated **X**'**y** refers to that part of **X**'**y**, which is orthogonal to the earlier scores $\mathbf{t}_1, .., \mathbf{t}_{h-1}$. The **X**'**X** matrix was deflated similarly. The deflation of **X**'**X** requires the regression of **X** onto the scores **T**, *i.e.* $\mathbf{X} = \mathbf{Tp} +$

f, where **p** are the loadings from this regression, and **f** are the residuals. We used the SIMPLS algorithm [14], which for a single trait **y** vector becomes http://www.statsoft.com/textbook/stpls.html:

1. Phenotypic values and marker genotype data were pre-treated and standardized in the same way as described in section 2.2.1 for PCR.

2. set $\mathbf{a}_1 = \mathbf{X}'\mathbf{y}$; $\mathbf{M}_1 = \mathbf{X}'\mathbf{X}$; and $\mathbf{C}_1 = \mathbf{I}$, then perform steps 3–8 for h = 1, ..., *c*:

3. $\mathbf{w}_h = \mathbf{a}_h/\text{sqrt}(\mathbf{a}_h'\mathbf{M}_h\mathbf{a}_h)$, which are the weights for the **X** columns to obtain $\mathbf{t}_h = \mathbf{X}\mathbf{w}_h$. The $\mathbf{w}_h$ are stored as columns in **W**.

4. $\mathbf{p}_h = \mathbf{M}_h\mathbf{w}_h$, which is the regression of **X** on $\mathbf{t}_h$. The $\mathbf{p}_h$ are stored as columns in **P**.

5. $\mathbf{q}_h = \mathbf{a}_h'\mathbf{w}_h$, which is the regression of **y** on $\mathbf{t}_h$. Since **y** is a single trait, $\mathbf{q}_h$ is a scalar and is stored in the column vector **q**.

6. $\mathbf{v}_h = \mathbf{C}_h\mathbf{p}_h$, standardised to have Euclidean length 1. The $\mathbf{v}_h$ is that part of $\mathbf{p}_h$, which is orthogonal to the earlier $\mathbf{p}_1$, .., $\mathbf{p}_{h-1}$ vectors.

7. $\mathbf{C}_{h+1} = \mathbf{C}_h - \mathbf{v}_h\mathbf{v}_h'$, which spans the space orthogonal to the $\mathbf{p}_1$, .., $\mathbf{p}_h$ vectors.

8. $\mathbf{a}_{h+1} = \mathbf{C}_{h+1}\mathbf{a}_h$, which deflates the $\mathbf{a}_h$ vector; and $\mathbf{M}_{h+1} = \mathbf{M}_h - \mathbf{p}_h\mathbf{p}_h'$, which deflates the $\mathbf{M}_h$ matrix. Return to step 3.

The regression coefficients of PLS regression then become $\mathbf{b}_{PLSR} = \mathbf{W}\mathbf{q}$. The correlation between TBV and EBV was calculated for *c* = 1, 2, 3, 4, 5, 7, 9, 12, 15 and 20 fitted latent variables. The number of latent variables, *c*, that maximised this correlation was used for each density.

'BayesB'

The 'BayesB' algorithm is described in detail and was used in earlier papers [4,7]. The 'BayesB' model was used to estimate marker effects and is briefly described as $\mathbf{y} = \mu\mathbf{1}_p + \Sigma_i\mathbf{Z}_i\mathbf{g}_i + \mathbf{e}$, where **y** is the vector of phenotypes, $\mathbf{1}_p$ is a vector of *p* ones, Σ_i is the summation over all markers, $\mathbf{Z}_i$ is a design matrix for the *i*'th marker, $\mathbf{g}_i$ is the vector of marker effects and **e** is the error. The variance of the marker effects (σ^2_{gi}) was estimated for every marker using a relevant prior distribution, which was a mixture of an inverted chi-squared distribution and a discrete probability mass at $\sigma^2_{gi} = 0$. A Metropolis-Hastings algorithm was used to sample σ^2_{gi} from its distribution conditional on $\mathbf{y}^*$, $p(\sigma^2_{gi} \mid \mathbf{y}^*)$, where $\mathbf{y}^*$ denotes the data **y** corrected for the mean and all other genetic effects except the marker effect ($\mathbf{g}_i$) [15]. Given σ^2_{gi}, marker effects, $\mathbf{g}_i$ was sampled from a Normal distribution as prior and using Gibbs sampling [16]. Estimated marker effects using 'BayesB' together with marker genotype of the animal was used to predict the breeding values, as explained in section 2.3.

Prediction of breeding values and statistics

Breeding values for the *n* animals in generation *t* = 1002 were estimated using the SNP marker information and the phenotypes in generation *t* = 1001, and compared to the true breeding values (TBV) in generation *t* = 1002. The EBV of animal *j* for PLSR and PCR were obtained from:

$$\text{EBV}_j = \mathbf{X}_j\mathbf{b}_a \text{ for } j = 1...n$$

where $\mathbf{X}_j$ denotes the *j*'th row of the **X** matrix corresponding to the set of genotypes for animal *j*, $\mathbf{b}_a$ is the regression coefficient vector of method *a*, where *a* denotes PCR or PLSR, and is estimated from the data in generation *t* = 1001. For 'BayesB' the EBVs were calculated from:

$$\text{EBV}_j = \sum_{i=1}^{m} Z_{i(j)}g_i \quad \text{for} \quad j = 1...n$$

where $\mathbf{Z}_{i(j)}$ denotes the row of the $\mathbf{Z}_i$ matrix corresponding to the genotype of animal *j* at locus *i*, and $\hat{\mathbf{g}}_i$ is the estimate of the marker effects for locus *i*, estimated in generation *t* = 1001.

TBV were linearly regressed on EBV, where the regression coefficient reflects the bias of the breeding value estimates (a regression coefficient of one denotes unbiased estimates), and the correlation coefficient reflects the accuracy of predicting the breeding values.

Results

Number of principal components with PCR

Figures 2 and 3 show the correlation of TBV with EBV and the regression of TBV on EBV as a function of the number of principal components for PCR. For the three lowest marker densities, $1N_e/M$, $2N_e/M$ and $4N_e/M$, the correlation reached a maximum when 250 principal components were fitted. For the $8N_e/M$ marker density, the correlation reached a maximum when 350 principal components were fitted. After reaching the highest correlation between TBV and EBV, the correlation coefficient between TBV and EBV was approximately maintained until dropping more steeply when the number of principal components exceeded 400 (Fig. 2). The regression coefficient decreased almost linearly, and hence the bias increased, as more principal components were fitted (Fig. 3). In the following tables and comparisons, the results from fitting 250 principal components were chosen for $1N_e/M$, $2N_e/M$ and $4N_e/M$ marker density schemes, while 350 principal

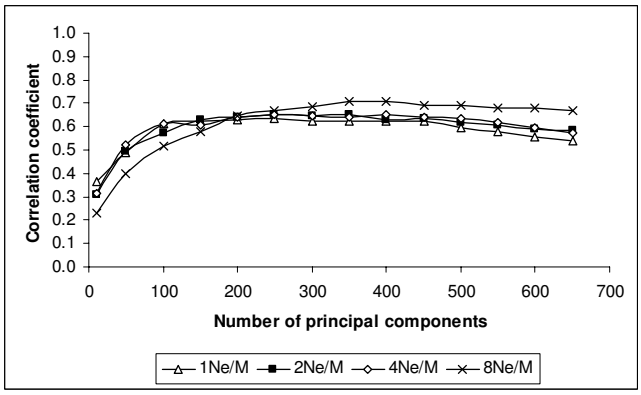

Figure 2
Correlation coefficient between TBV and EBV using principal component regression (PCR) for different marker density schemes (1, 2, 4 and $8N_e/M$) when the number of principal components was varied.

components were chosen for the $8N_e/M$ marker density scheme, since this achieved the highest correlation between TBV and EBV with PCR.

Number of latent variables with PLSR

The correlation coefficient between TBV and EBV and the regression coefficient of TBV on EBV resulting from varying the number of latent variables from 1 to 20 are shown in Figures 4 and 5, respectively. Starting with one latent variable, the correlation coefficient between TBV and EBV increased until it reached a maximum between 2–4 latent variables (Fig. 4). The regression of TBV on EBV was close to 1 when only one latent variable was fitted, and dropped rapidly as more latent variables were added to the model

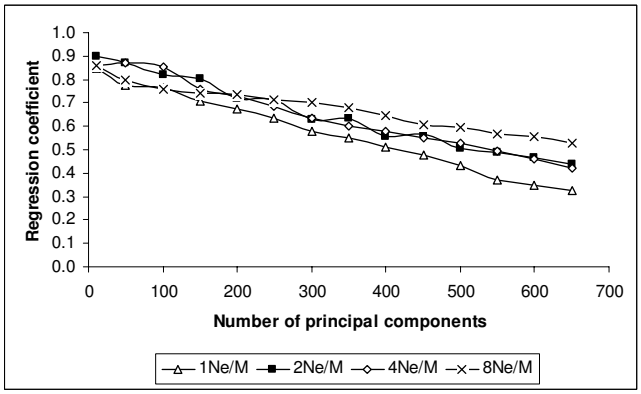

Figure 3
Regression coefficient of TBV on EBV using principal component regression (PCR) for different marker density schemes (1, 2, 4 and $8N_e/M$) when the number of principal components was varied.

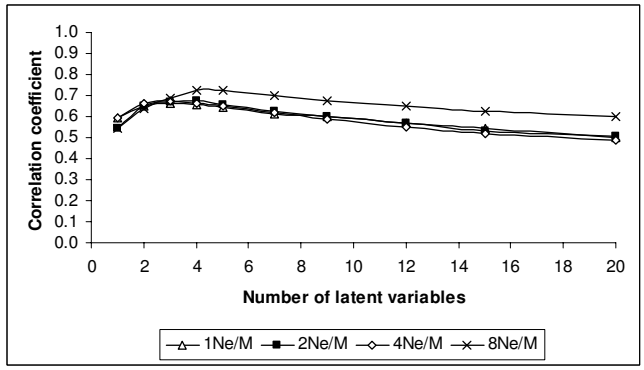

Figure 4
Correlation coefficient between TBV and EBV using partial least square regression (PLSR) for different marker density schemes (1, 2, 4 and $8N_e/M$) when the number of latent variables was varied.

(Fig. 5). In the following tables and comparisons the results from fitting two latent variables for the marker densities $1N_e/M$, $2N_e/M$ and $4N_e/M$ were chosen, while four latent variables were chosen for the $8N_e/M$ marker density scheme, since this achieved the highest correlation between TBV and EBV with PLSR.

Correlation

The correlation coefficients between TBV and EBV for the different marker densities and estimation methods together with their standard error are given in Table 1. The accuracy of estimating the breeding values increased as the density of the markers increased, as expected, since more information was available when more markers were fitted to the model. For PLSR, the correlation coefficient between TBV and EBV for the four densities (1, 2, 4 and $8N_e/M$) was 0.611, 0.655, 0.670 and 0.681, respectively.

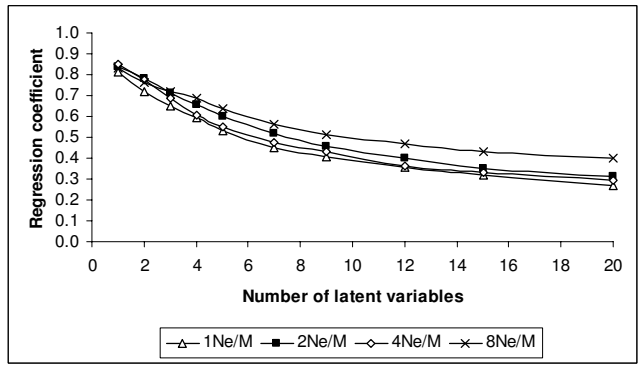

Figure 5
Regression coefficient of TBV on EBV using partial least square regression (PLSR) for different marker density schemes (1, 2, 4 and $8N_e/M$) when the number of latent variables was varied.

Table 1: The mean correlation ($r_{TBV;\ EBV}$) between TBV and EBV using principal component regression (PCR), partial least square regression (PLSR) and the 'BayesB' method for different marker densities, averaged over 20 replicates

	PCR	PLSR	'BayesB'
Marker density	$r_{TBV;\ EBV}$ ± s.e	$r_{TBV;\ EBV}$ ± s.e	$r_{TBV;\ EBV}$ ± s.e
$1N_e/M$	0.604 ± 0.012	0.611 ± 0.012	0.690 ± 0.036
$2N_e/M$	0.639 ± 0.012	0.655 ± 0.012	0.790 ± 0.036
$4N_e/M$	0.645 ± 0.012	0.670 ± 0.012	0.841 ± 0.036
$8N_e/M$	0.665 ± 0.012	0.681 ± 0.012	0.860 ± 0.036

This was marginally greater than using PCR, which varied in a similar fashion from 0.604 to 0.665. Compared within densities the differences between PCR and PLSR were not significant, but across all densities PLSR gave a higher correlation than PCR by 0.016 (s.e = 0.008). The correlation coefficient between TBV and EBV for 'BayesB' was 8% greater than PLSR for the lowest marker density, and 18% greater for the highest marker density. Hence, the gap in accuracy between PLSR/PCR and 'BayesB' increased as the marker density increased.

Regression

The regression coefficients of TBV on EBV are summarized in Table 2. The most evident result is that this regression was higher for 'BayesB' compared to the regression methods, PLSR and PCR: the mean coefficient for 'BayesB' was > 0.87 for all marker densities, but was always < 0.76 for the regression methods, and this difference was large compared to the standard errors obtained. For 'BayesB' there was statistical evidence for a trend towards regression coefficients increasing towards 1 as marker densities increased. For the two regression methods, the pattern was more complex. More principal components and latent variables were fitted to optimise the correlations shown in Table 1 for $8N_e/M$ than in scenarios with lower marker density, *i.e.* 350 principal components and four latent variables were used for $8N_e/M$, while 250 principal components and two latent variables were used for PCR and PLSR respectively for the lower marker density schemes. Figures 3 and 5 show clearly that the regression coefficient decreases as the number of principal components or latent variables increase for both methods. With this caveat, at low densities (1, 2 and 4 N_e/M) it appeared that the PLSR method resulted in greater regression coefficients than PCR (a difference of 0.07 with s.e = 0.01 over these densities), but this was reversed in favour of PCR (0.036 with s.e = 0.018) at $8N_e/M$. The regression coefficients for PCR appeared more stable compared to PLSR and exhibited a trend to greater regression coefficients as marker density increased.

Computer time

Compared to the 'BayesB' method, the presented multivariate regression methods used much less computational time. The computer time for estimating the marker effects using the PCR, PLSR and 'BayesB' is presented in Table 3. The machine was an HP AlphaServer GS1280 with eight processors (EV7), of which only one processor was used at a time. PLSR used about 3 min per replicate to compute the marker effects for all marker densities, while PCR used somewhat longer time to calculate the marker effects, especially for higher marker densities, and the gap in computation time between PLSR and PCR increased as the marker density increased. However, the computation time for PLSR/PCR was very much reduced compared to the 'BayesB': *e.g.* 'BayesB' used approximately 200 minutes to compute the marker effects for the lowest marker density, which was approximately 65 times longer than PLSR/PCR, and the computer time increased rapidly as the marker density increased (Table 3).

Discussion

Two multivariate regression methods that reduce the dimensionality of the marker data were compared to a Bayesian method for the prediction of genome-wide breeding values based on SNP marker information and phenotypic records. In general, our results showed that it was possible to predict breeding values in our simulated genome using both multivariate regression methods, but the correlation between TBV and EBV were both reduced compared to those of 'BayesB'. The correlation between TBV and EBV increased as the marker density increased,

Table 2: The mean regression coefficient ($b_{TBV;\ EBV}$) between TBV on EBV using principal component regression (PCR), partial least square regression (PLSR) and the 'BayesB' method for different marker densities, averaged over 20 replicates.

	PCR	PLSR	'BayesB'
Marker density	$b_{TBV;\ EBV}$ ± s.e	$b_{TBV;\ EBV}$ ± s.e	$b_{TBV;\ EBV}$ ± s.e
$1N_e/M$	0.650 ± 0.012	0.758 ± 0.013	0.877 ± 0.013
$2N_e/M$	0.683 ± 0.012	0.725 ± 0.013	0.879 ± 0.013
$4N_e/M$	0.695 ± 0.012	0.754 ± 0.013	0.943 ± 0.013
$8N_e/M$	0.691 ± 0.012	0.655 ± 0.013	0.923 ± 0.013

The standard errors (s.e) shown are derived from the pooled variance between replicates within each evaluation method.

Table 3: Computation time for estimating the marker effects using principal component regression (PCR), partial least square regression (PLSR) and the 'BayesB' method

Marker density	PCR	PLSR	'BayesB'
$1N_e/M$	~3 min	~3 min	~200 min
$2N_e/M$	~15 min	~3 min	~700 min
$4N_e/M$	~30 min	~3 min	~1600 min
$8N_e/M$	~60 min	~3 min	> 2800 min

because more information was available for predicting QTL genotypes, but most notably for 'BayesB'. The correlation is the accuracy of predicting EBV, whilst the regression indicates bias, and these correspondences will be used throughout the rest of the discussion. Hence, the results indicate that the regression methods deliver a lower accuracy and greater bias in predicting breeding values than 'BayesB', and are less responsive to the addition of further marker information.

The greater responsiveness to marker density of 'BayesB' was marked. For PLSR and PCR, the accuracy increased by 7% and 6% respectively from the lowest marker density ($1N_e/M$) to the highest marker density ($8N_e/M$), whilst in contrast 'BayesB' was 17% more accurate for the highest marker density compared to the lowest density. Hence, the gap in accuracy between PLSR/PCR and 'BayesB' increased as the marker density increased. From this result, it seems that the use of relevant prior information, as in the 'BayesB' method, was more valuable as the marker density increased.

Whilst the accuracy of prediction may be the primary parameter of interest, the regression of the TBV on EBV is relevant since it determines the bias in predicting genetic progress. One possible consequence is that this will contribute to decreasing the accuracy in predicting breeding values if the population used for providing estimates of breeding values spans more than a single generation of selection. In this attribute, as in accuracy, the advantage appears to lie in 'BayesB', with regression coefficients both closer to one than PLSR and PCR and increasing as marker density increased. Although, these biases may be corrected for by scaling the EBV such that Var(EBV) = Cov(EBV, TBV), and thus are not a major hindrance for the use of PLSR or PCR. The regression methods had increased bias as density increased because more principal components or latent variables were required to optimise the accuracy. Any use of PLSR or PCR would require optimisation on the number of principal components or latent variables, perhaps through cross-validation for each practical data set, although both accuracy and bias will depend on the number of phenotypes.

The main advantages using PLSR and PCR compared to the 'BayesB' method were the computing time and avoidance of the assumptions about prior distribution of marker effects made in the 'BayesB' model. PLSR and PCR were computationally much faster and simpler compared to the 'BayesB' method, *e.g.* the computation time for estimating the marker effects using PLSR was approximately 65 times faster than 'BayesB' for the lowest marker density. The gap in computation time, hence the computational costs, were increased for higher marker density. For example, the expected linkage disequilibrium (LD) for the same recombination rate will be reduced by doubling the effective population size N_e. Hence, assuming the accuracy is primarily determined by the amount of LD, then a doubling of the number of markers is needed to achieve the same LD, a finding supported by [7]. Doubling the number of markers will double or triple the computation time needed, which is especially time consuming for 'BayesB'. Compared to PLSR and PCR, 'BayesB' has a greater potential for exploiting parallel computing, which was not used in this study, therefore the relative computational benefits of PLSR and PCR will diminish as parallel processing becomes cheaper and more common. This parallel computing implementation of BayesB will be highly needed because the number of markers is expected to increase for most species to 50 – 500 thousand.

Meuwissen *et al.* [4] used microsatellites at 1 N_e/M density to compare least square regression after screening for significant QTL, BLUP and 'BayesB' for predicting genome-wide breeding values, and found accuracies of 0.318, 0.732 and 0.848, respectively. Solberg *et al.* [7] determined that SNP densities of 2- to 3-fold greater densities were required to achieve comparable accuracies. Therefore, an appropriate comparison may be made with the results of [4] with SNP at a density of $4N_e/M$ in our study. For this density, PLSR and PCR had accuracies of 0.670 and 0.645. Hence, these results indicate that the Bayesian method 'BayesB' gave the highest accuracy, followed by BLUP, PLSR and PCR, and least square analysis combined with screening had the lowest accuracy.

A somewhat high heritability was used in this simulation study, therefore a comparison of the three methods BayesB, PLSR and PCR were additionally evaluated for a lower heritability of 0.25 (Table 4). For the highest marker density, the selection accuracy was reduced by 7% for the BayesB method, and 16% for the two regression methods. For the lowest marker density, the selection accuracy was reduced by 14% for the regression methods and 12% for the BayesB method. No significant differences were observed between the PLSR and PCR. Even if the selection accuracy was reduced in all cases, the same "ranking" of the methods remain, namely, BayesB performed better than PLSR and PCR.

Table 4: Comparison of the three methods for the lowest (1Ne/M) and the highest (8Ne/M) marker density, when the heritability was 0.25

	PCR	PLSR	'BayesB'
Marker density	$r_{TBV;\ EBV} \pm$ s.e	$r_{TBV;\ EBV} \pm$ s.e	$r_{TBV;\ EBV} \pm$ s.e
$1N_e/M$	0.452 ± 0.009	0.465 ± 0.011	0.566 ± 0.018
$8N_e/M$	0.510 ± 0.012	0.504 ± 0.014	0.793 ± 0.018

A conclusion from this study is that if some relevant information is known *a priori* then methods that utilize relevant prior information will be more accurate. The 'BayesB' method assumed a mixture of distributions of an inverted chi-square with a discrete probability mass at zero as the relevant prior distribution of marker effects, to model an increase in the number of markers with an effect of zero. The simulated QTL effects followed a gamma distribution with a shape parameter of 1.66 and a scale parameter of 0.4 [12] with equal probability of positive or negative effects. In practice, we do not know the exact distribution of the QTL effects. Although the distribution used for simulating the QTL effects and that used for analysing the data did not agree exactly, 'BayesB' approximates the prior distribution of the QTL effects better than the regression methods. From a Bayesian perspective, PLSR and PCR might be viewed as representing a limiting form where the prior distribution for regression coefficients is normally distributed with an increasingly large variance. This closer correspondence between the prior used for evaluation and the simulated distribution of QTL perhaps explains in part the higher accuracies obtained with 'BayesB'.

PLSR and PCR give an alternative solution to 'BayesB' to estimate marker effects. They provide a rapid analysis of large amounts of data to obtain EBVs from high-density markers. The only assumptions are the additivity of marker effects, and that few linear combinations of markers can explain most variability in the data. However, whilst this simulation study showed that reducing the dimensionality of the data gave a reasonably high accuracy of selection, the accuracy was less than that obtained from 'BayesB', and this difference increased with increasing marker density. To obtain full benefits of genome-wide selection, use of relevant *a priori* information about the distribution of the QTL effects is preferable, since genotyping costs are very high relative to computational costs. These relevant prior distributions need to be obtained by acquiring greater knowledge of the genomic architecture.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TRS simulated the datasets, carried out the analysis and drafted the manuscript. TM helped to carry out the study and drafting the manuscript. All authors have read and approved the final manuscript.

References

1. Gianola D, Fernando RL, Stella A: **Genomic-assisted prediction of genetic value with semiparametric procedures.** *Genetics* 2006, **173:**1761-1776.
2. Gianola D, Perez-Enciso M, Toro MA: **On marker-assisted prediction of genetic value: beyond the ridge.** *Genetics* 2003, **163:**365-374.
3. Habier D, Fernando RL, Dekkers JCM: **The impact of genetic relationship information on genome-assisted breeding values.** *Genetics* 2007, **177:**2389-2397.
4. Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157:**1819-1829.
5. Muir WM: **Genomic selection: a break through for application of marker assisted selection to traits of low heritability, promise and concerns.** *58th EAAP; Dublin, Ireland* 2007.
6. Schaeffer LR: **Strategy for applying genome-wide selection in dairy cattle.** *J Anim Breed Genet* 2006, **123:**218-223.
7. Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE: **Genomic selection using different marker types and densities.** *J Anim Sci* 2008, **86:**2447-2454. (Published online Apr 11, 2008, doi:10.2527/jas.2007-0010)
8. Martens H, Næs T: *Multivariate calibration* John Wiley & Sons Ltd; 1991. ISBN 0-471 93047-4
9. Wold H: **Estimation of principal components and related models by iterative least squares.** In *Multivariate analysis* Edited by: Krishnaiah PR. New York: Academic Press; 1966.
10. Pinto LFB, Packer IU, De Melo CMR, Ledur MC, Coutinho LL: **Principal component analysis applied to performance and carcass traits in the chicken.** *Anim Res* 2006, **55:**419-425.
11. Sölkner J, Tier B, Crump R, Moser G, Thomson P, Raadsma H: **A comparison of different regression methods for genomic-assisted prediction of genetic values in dairy cattle.** *58th EAAP; Dublin, Ireland* 2007.
12. Hayes BJ, Goddard ME: **The distribution of the effects of genes affecting quantitative traits in livestock.** *Genet Sel Evol* 2001, **33:**209-229.
13. Press WH, Teukolsky SA, Vetterling WT, Flanery BP: *Numerical recipes in Fortran 77: The art of scientific computing* Second edition. 2003. ISBN 0-521-43064-X
14. de Jong S: **SIMPLS: an alternative approach to partial least square regression.** *J Chemometrics* 1993, **12:**41-54.
15. Gilks WR, Richardson S, Spiegelhalter DJ: *Markov Chain Monte Carlo in practice* Chapman & Hall/CRC; 1996. ISBN 0-412-05551-1
16. Sørensen D, Gianola D: *Likelihood, bayesian, and MCMC methods in quantitative genetics* Springer; 2002. ISBN 0-387-95440-6